CODA-19: Reliably Annotating Research Aspects on 10,000+ CORD-19 Abstracts Using a Non-Expert Crowd

Ting-Hao (Kenneth) Huang[1], **Chieh-Yang Huang**[1] **Chien-Kuang Cornelia Ding**[2], **Yen-Chia Hsu**[3], **C. Lee Giles**[1]
[1]Pennsylvania State University, University Park, PA, USA
{txh710, chiehyang, clg20}@psu.edu
[2]University of California, San Francisco, CA, USA. Cornelia.Ding@ucsf.edu
[3]Carnegie Mellon University, Pittsburgh, PA, USA. yenchiah@andrew.cmu.edu

Abstract

This paper introduces **CODA-19**[1], a human-annotated dataset that codes the **Background, Purpose, Method, Finding/Contribution, and Other** sections of 10,966 English abstracts in the COVID-19 Open Research Dataset. CODA-19 was created by 248 crowd workers from Amazon Mechanical Turk within 10 days, achieving a label quality comparable to that of experts. Each abstract was annotated by nine different workers, and the final labels were obtained by majority vote. The inter-annotator agreement (Cohen's kappa) between the crowd and the biomedical expert (0.741) is comparable to inter-expert agreement (0.788). CODA-19's labels have an accuracy of 82.2% when compared to the biomedical expert's labels, while the accuracy between experts was 85.0%. Reliable human annotations help scientists to understand the rapidly accelerating coronavirus literature and also serve as the battery of AI/NLP research, but obtaining expert annotations can be slow. We demonstrated that a non-expert crowd can be rapidly employed at scale to join the fight against COVID-19.

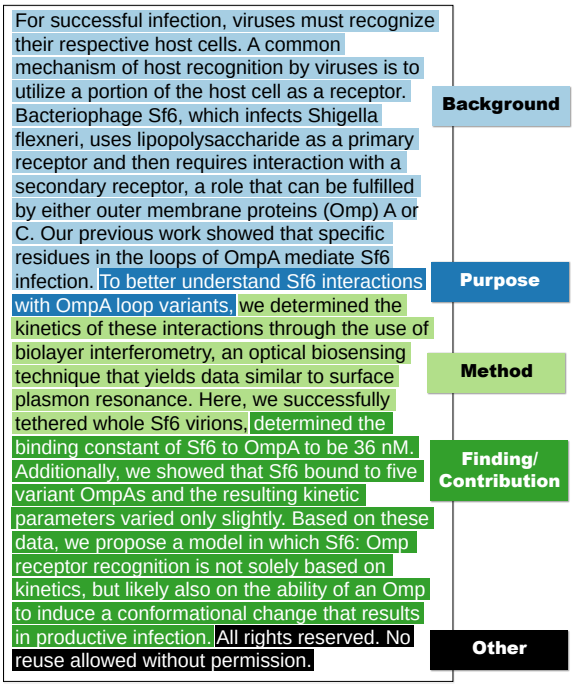

Figure 1: An example of the final crowd annotation for the abstract of (Hubbs et al., 2019).

1 Introduction

Just as COVID-19 is rapidly spreading worldwide, the rapid acceleration in new coronavirus literature makes it hard to keep up with. Researchers have thus teamed up with the White House to release the COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020), containing over 59,000 related scholarly articles (as of May 1, 2020). The Open Research Dataset Challenge has also been launched on Kaggle to encourage researchers to use cutting-edge techniques to gain new insights from these papers. However, it often requires large-scale human annotations for automated language understanding, relation extraction, and question answering to reach good performance levels. Producing such annotations for thousands of papers can be a prolonged process if we only employ expert annotators, whose availability is limited.

Data sparsity is one of the challenges for text mining in the biomedical domain because text annotations on scholarly articles were mainly produced by small groups of experts. For example, two researchers manually created the ACL RD-TEC 2.0, a dataset that contains 300 scientific abstracts (QasemiZadeh and Schumann, 2016); a group of annotators "with rich experience in biomedical content curation" created MedMentions, a corpus containing 4,000 abstracts (Mohan and Li, 2019); and several datasets used in biomed-

[1]COVID-19 Research Aspect Dataset (CODA-19): https://github.com/windx0303/CODA-19

ical NLP shared tasks were manually created by the organizers and/or their students, such as the ScienceIE in SemEval'17 (Augenstein et al., 2017) and Relation Extraction in SemEval'18 (Gábor et al., 2018). Obtaining expert annotations can be too slow to respond to COVID-19, so we explore an alternative approach: using **non-expert crowds**, such as workers on Amazon Mechanical Turk (MTurk), to produce high-quality, useful annotations for thousands of scientific papers.

This paper introduces **CODA-19**, the **COVID-19 Research Aspect Dataset**, presenting the first outcome of our exploration in using non-expert crowds for large-scale scholarly article annotation. CODA-19 contains 10,966 abstracts randomly selected from CORD-19. Each abstract was segmented into sentences, which were further divided into one or more shorter text fragments. All 168,286 text fragments in CODA-19 were labeled with a "research aspect," *i.e.*, **Background, Purpose, Method, Finding/Contribution, or Other**. This annotation scheme was adapted from SOLVENT (Chan et al., 2018), with minor changes.

In our project, 248 crowd workers from MTurk were recruited and annotated the whole CODA-19 within ten days.[2] Each abstract was annotated by nine different workers. We aggregated the crowd labels for each text segment using majority voting.

The resulting crowd labels had a label accuracy of 82% when compared against the expert labels on 129 abstracts. The inter-annotator agreement (Cohen's kappa) was 0.741 between the crowd labels and the expert labels, while it was 0.788 between two experts. We also established several classification baselines, showing the feasibility of automating such annotation tasks.

2 Annotation Scheme

CODA-19 uses a five-class annotation scheme to denote research aspects in scientific articles: **Background, Purpose, Method, Finding/Contribution, or Other**. Table 1 shows the full annotation guidelines we developed to instruct workers. We updated and expanded this guideline daily during the annotation process to address workers' questions and feedback.

This scheme was adapted from SOLVENT (Chan et al., 2018), with three changes. First, we added an "Other" category. Articles in CORD-19 are broad and diverse (Colavizza et al., 2020), so it is unrealistic to govern all cases with only four categories. We are also aware that CORD-19's data came with occasional formatting or segmenting errors. These cases were also to be put into the "Other" category. Second, we replaced the "Mechanism" category with "Method." Chan *et al.* created SOLVENT with the aim of discovering the analogies between research papers at scale. Our goal was to better understand the contribution of each paper, so we decided to use a more general word, "Method," to include the research methods and procedures that cannot be characterized as "Mechanisms." Also, biomedical literature widely used the word "mechanism," which could also be confusing to workers. Third, we modified the name "Finding" to "Finding/Contribution" to allow broader contributions that are not usually viewed as "findings." Our scheme is also similar to that of DISA (Huang and Chen, 2017), which has an additional "Conclusion" category.

We selected this scheme because it balances the richness of information and the difficulty level for workers to annotate. We are aware of the long history of research (Kilicoglu, 2018) on composing structured abstracts (Hartley, 2004), identifying argumentative zones (Teufel et al., 1999; Mizuta et al., 2006; Liakata et al., 2010), analyzing scientific discourse (de Waard and Maat, 2012; Dasigi et al., 2017; Banerjee et al., 2020), supporting paper writing (Wang et al., 2019), and representing papers to reduce information overload (de Waard et al., 2009). However, most of these schemes assumed expert annotators rather than crowd workers. We eventually narrowed our focus down to two annotation schemes: SOLVENT and the "Information Type" (Focus, Polarity, Certainty, Evidence, Trend) proposed by Wilbur *et al.* (2006). SOLVENT is easier to annotate and has been tested with workers from MTurk and Upwork, while Wilbur's scheme is informative and specialized for biomedical articles. We implemented annotation interfaces for both schemes and launched a few tasks on MTurk for testing. Workers accomplished the SOLVENT tasks much faster with reasonable label accuracy, while only a few workers accomplished the Information Type annotation task. Therefore, we decided to adapt the SOLVENT scheme.

[2] From April 19, 2020 to April 29, 2020, including the time for worker training and post-task survey.

Aspect	Annotation Guideline
Background	"Background" text segments answer one or more of these questions: • Why is this problem important? • What relevant works have been created before? • What is still missing in the previous works? • What are the high-level research questions? • How might this help other research or researchers?
Purpose	"Purpose" text segments answer one or more of these questions: • What specific things do the researchers want to do? • What specific knowledge do the researchers want to gain? • What specific hypothesis do the researchers want to test?
Method	"Method" text segments answer one or more of these questions: • How did the researchers do the work or find what they sought? • What are the procedures and steps of the research?
Finding/ Contribution	"Finding/Contribution" text segments answer one or more of these questions: • What did the researchers find out? • Did the proposed methods work? • Did the thing behave as the researchers expected?
Other	• Text segments that do *not* fit into any of the four categories above. • Text segments that are *not* part of the article. • Text segments that are *not* in English. • Text segments that contain *only* reference marks (*e.g.*, "[1,2,3,4,5") or dates (*e.g.*, "April 20, 2008"). • Captions for figures and tables (*e.g.* "Figure 1: Experimental Result of ...") • Formatting errors. • Text segments the annotator does not know or is not sure about.

Table 1: CODA-19's annotation guideline for crowd workers.

3 CODA-19 Dataset Construction

CODA-19 has 10,966 abstracts that contain a total of 2,703,174 tokens and 103,978 sentences, which were divided into 168,286 segments. The data is released as a 80/10/10 train/dev/test split.

3.1 Data Preparation

We used Stanford CoreNLP (Manning et al., 2014) to tokenize and segment sentences for all the abstracts in CORD-19. We further used comma (,), semicolon (;), and period (.) to split each sentence into shorter fragments, where a fragment has no fewer than six tokens (including punctuation marks) and has no orphan parentheses.

As of April 15, 2020, 29,306 articles in CORD-19 had a non-empty abstract. An average abstract had 9.73 sentences (SD = 8.44), which were further divided into 15.75 text segments (SD = 13.26). Each abstract had 252.36 tokens (SD = 192.89) on average. We filtered out the 538 (1.84%) abstracts with only one sentence because many of them had formatting errors. We also removed the 145 (0.49%) abstracts that had more than 1,200 tokens to keep the working time for each task under five minutes (see Section 3.3). We randomly selected 11,000 abstracts from the remaining data for annotation. During the annotation process, workers informed us that a few articles were not in English. We identified these automatically using `langdetect`[3] and excluded them.

3.2 Interface Design

Figure 2 shows the worker interface, which we designed to guide workers to read and label all the text segments in an abstract. The interface showed the instruction on the top (Figure 2a) and presented the task in three steps: In Step 1, the worker was instructed to spend ten seconds to take a quick glance at the abstract. The goal was to get a high-level sense of the topic rather than to fully understand the abstract. In Step 2, we showed the main annotation interface (Figure 2b), where the worker can go through each text segment and select the most appropriate category for each segment one by one. In Step 3, the worker can review the labeled text segments (Figure 2c) and go back to Step 2 to fix any problems.

3.3 Annotation Procedure

Worker Recruiting We first created a qualification Human Intelligence Task (HIT) to recruit workers on MTurk ($1/HIT). The workers needed to watch a five-minute video to learn the scheme, go

[3] langdetect: https://github.com/Mimino666/langdetect

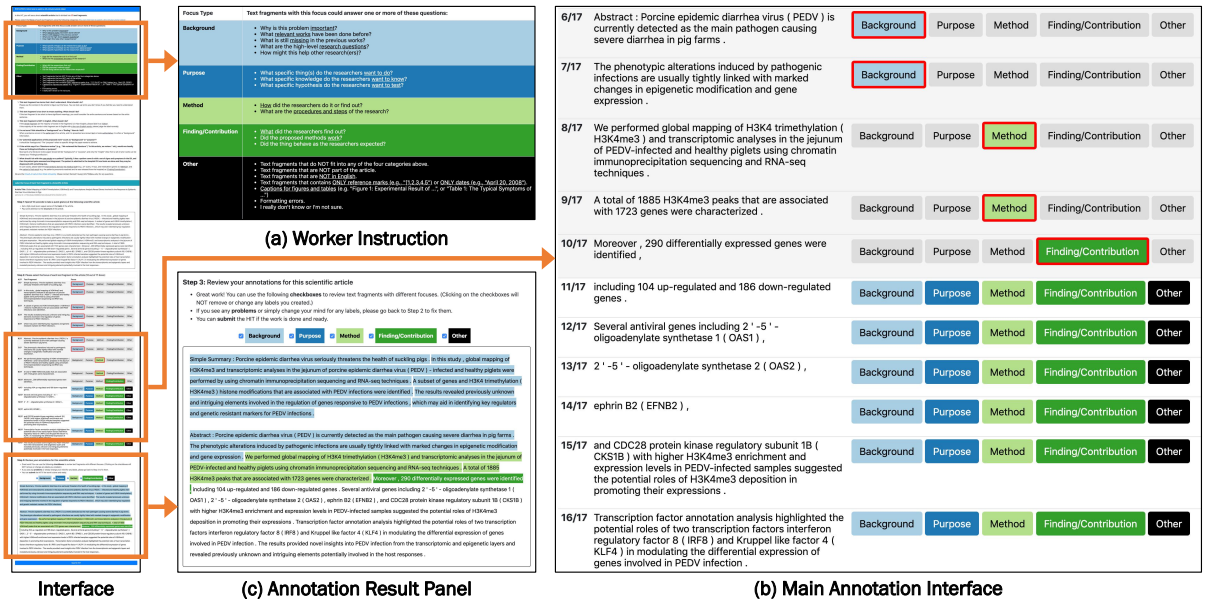

Figure 2: The worker interface used to construct CODA-19.

through an interactive tutorial to learn the interface, and sign a consent form to obtain the qualification. We granted custom qualifications to 400 workers who accomplished the qualification HIT. Only the workers with this qualification could do our tasks. Four built-in MTurk qualifications were also used: Locale (US Only), HIT Approval Rate (≥98%), Number of Approved HITs (≥3000), and the Adult Content Qualification.

Posting Tasks in Smaller Batches We divided 11,000 abstracts into smaller batches, where each batch has no more than 1,000 abstracts. Each abstract forms a single HIT. We recruited nine different workers through nine assignments to label each abstract. Our strategy was to post one batch at a time. When a batch was finished, we assessed its data quality, sent feedback to workers to guide them, or blocked workers who constantly had low accuracy before proceeding with the next batch.

Worker Wage and Total Cost We aimed to pay an hourly wage of $10. The working time of an abstract was estimated by the average reading speed of English native speakers, *i.e.*, 200-300 words per minute (Siegenthaler et al., 2012). For an abstract, we rounded up $(\#token/250)$ to an integer as the estimated working time in minutes and paid $(\$0.05 + \text{Estimated Working Minutes} \times \$0.17)$ for it. As a result, 59.49% of our HITs were priced at $0.22, 36.41% were at $0.39, 2.74% were at $0.56, 0.81% were at $0.73, and 0.55% were at $0.90. We posted nine assignments per HIT. Adding the 20% MTurk fee, coding each abstract (using nine workers) cost $3.21 on average.

3.4 Label Aggregation

The final labels in CODA-19 were obtained by majority voting over crowd labels, excluding the labels from blocked workers. For each batch of HITs, we manually examined the labels from workers who frequently disagreed with the majority-voted labels (Section 3.3). If a worker had abnormally low accuracy or was apparently spamming, we retracted the worker's qualification to prevent him/her from taking future tasks. We excluded the labels from these removed workers when aggregating the final labels. Note that there can be ties when two or more aspects received the same highest number of votes (*e.g.,* 4/4/1 or 3/3/3). We resolved ties by using the following tiebreakers, in order: Finding, Method, Purpose, Background, Other.

4 Data Quality Assessment

We worked with a biomedical expert and a computer scientist to assess label quality; both experts are co-authors of this paper. The biomedical expert (the "Bio" Expert in Table 2) is an MD and also a PhD in Genetics and Genomics. She is now a resident physician in pathology at the University of California, San Francisco. The other expert (the "CS" Expert in Table 2) has a PhD in Computer Science and is currently a Project Scientist at Carnegie Mellon University.

Eval. Label	Gold Label	Background			Purpose			Method			Finding			Other			acc	kappa
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1		
Crowd	Bio	.827	.911	.867	.427	.662	.519	.783	.710	.744	.874	.838	.856	.986	.609	.753	.822	.741
Crowd	CS	.846	.883	.864	.700	.611	.653	.818	.633	.714	.800	.931	.860	.986	.619	.761	.821	.745
CS	Bio	.915	.966	.940	.421	.746	.538	.670	.785	.723	.958	.789	.865	.867	.852	.860	.850	.788

Table 2: Crowd performance using both Bio Expert and CS Expert as the gold standard. CODA-19's labels have an accuracy of 0.82 and a kappa of 0.74, when compared against two experts' labels. It is noteworthy that when we compared labels between two experts, the accuracy (0.850) and kappa (0.788) were only slightly higher.

Both experts annotated the same 129 abstracts randomly selected from CODA-19. The experts used the same interface as that of the workers (Figure 2). The inter-annotator agreement (Cohen's kappa) between the two experts was 0.788. Table 2 shows the aggregated crowd label's accuracy, along with the precision, recall, and F1-score of each class. CODA-19's labels have an accuracy of 0.82 and a kappa of 0.74 when compared against the two experts' labels. It is noteworthy that when we compared labels between the two experts, the accuracy (0.850) and kappa (0.788) were only slightly higher. The crowd workers performed best in labeling "Background" and "Finding," and they had nearly perfect precision for the "Other" category. Figure 3 shows the normalized confusion matrix for the aggregated crowd labels versus the biomedical expert's labels. Many "Purpose" segments were mislabeled as "Background," which might indicate more ambiguous cases between these two categories. During the annotation period, we received several emails from workers asking about the distinctions between these two aspects. For example, do "potential applications of the proposed work" count as "Background" or "Purpose"?

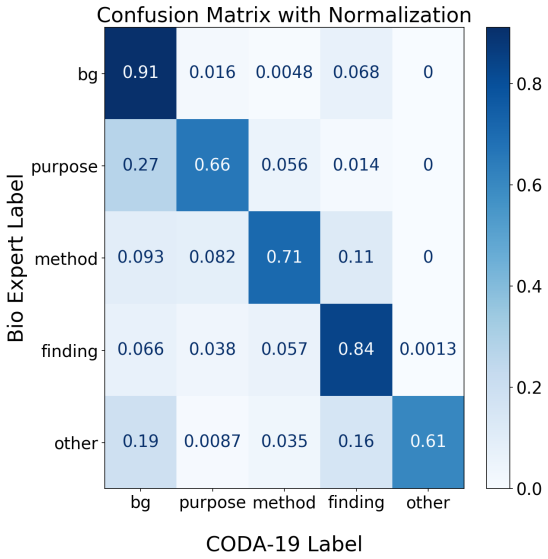

Figure 3: The normalized confusion matrix for the CODA-19 labels versus the biomedical expert's labels.

5 Classification Baselines

We further examined machines' capacity for annotating research aspects automatically. Six baseline models were implemented: Linear SVM, Random Forest, CNN, LSTM, BERT, and SciBERT.

Data Preprocessing The Tf-idf feature was used. We turned all words into lowercase and removed those with frequency lower than 5. The final tf-idf feature contained 16,775 dimensions. For deep-learning approaches, the vocabulary size was 16,135, where tokens with frequency lower than 5 were replaced by `<UNK>`. Sequences were padded with `<PAD>` if containing less than 60 tokens and were truncated if containing more than 60 tokens.

Models Machine-learning approaches were implemented using Scikit-learn (Pedregosa et al., 2011) and deep-learning approaches were implemented using PyTorch (Paszke et al., 2019). The following are the training setups.

- **Linear SVM:** We did a grid search for hyperparameters and found that $C = 1$, $tol = 0.001$, and *hinge loss* yielded the best results.
- **Random Forest:** With the grid search, 150 estimators yielded the best result.
- **CNN:** The classic CNN (Kim, 2014) was implemented. Three kernel sizes (3,4,5) were used, each with 100 filters. The word embedding size was 256. A dropout rate of 0.3 and L2 regularization with weight 10^{-6} were used when training. We used the Adam optimizer, with a learning rate of 0.00005. The model was trained for 50 epochs and the one with highest validation score was kept for testing.

Model	Background			Purpose			Method			Finding			Other			Accuracy
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
# Sample	5062			821			2140			6890			562			15475
SVM	.658	.703	.680	.621	.446	.519	.615	.495	.549	.697	.729	.712	.729	.699	.714	.672
RF	.671	.632	.651	.696	.365	.479	**.716**	.350	.471	.630	**.787**	.699	.674	.742	.706	.652
CNN	.649	.706	.676	.612	.512	.557	.596	.562	.579	.726	.702	.714	.743	.795	.768	.677
LSTM	.655	.706	.680	**.700**	.464	.558	.634	.508	.564	.700	.724	.711	.682	.770	.723	.676
BERT	.719	.759	.738	.585	**.639**	.611	.680	.612	.644	.777	.752	.764	.773	**.874**	.820	.733
SciBERT	**.733**	**.768**	**.750**	.616	.636	**.626**	.715	**.636**	**.673**	**.783**	.775	**.779**	**.794**	.852	**.822**	**.749**

Table 3: Baseline performance of automatic labeling using the crowd labels of CODA-19. SciBERT achieves highest accuracy of 0.749 and outperforms other models in every aspects.

- **LSTM:** We used 10 LSTM layers to encode the sequence. The encoded vector was then passed through a dense layer for classification. Word embedding size and LSTM hidden size were both 256. The rest of the hyper-parameter and training setting was the same as that of the CNN model.

- **BERT:** Hugging Face's implementation (Wolf et al., 2019) of the Pretrained BERT (Devlin et al., 2018) was used for fine-tuning. We fine-tuned the pretrained model with a learning rate of $3 * 10^{-7}$ for 50 epochs. Early stopping was used when no improvement occurred in the validation accuracy for five consecutive epochs. The model with the highest validation score was kept for testing.

- **SciBERT:** Hugging Face's implementation (Wolf et al., 2019) of the Pretrained SciBERT (Beltagy et al., 2019) was used for fine-tuning. The fine-tuning setting is the same as that of the BERT model.

Result Table 3 shows the results for the six baseline models: SciBERT preformed the best in overall accuracy. When looking at each aspect, all the models performed better in classifying "Background," "Finding," and "Other," while identifying "Purpose" and "Method" was more challenging.

6 What's Next?

One obvious future direction is to improve classification performance. We evaluated the automatic labels against the biomedical expert's labels, and the SciBERT model achieved an accuracy of 0.774 and a Cohen's kappa of 0.667, indicating some space for further improvement. Our baseline approaches did not use any contextual information nor domain knowledge. We expect that the classification performance can be further boosted, allowing researchers to label future papers automatically.

How can these annotations help search and information extraction? Several search engines have been quickly developed and deployed. These engines allow users to navigate CORD-19 more efficiently and could potentially support decision-making. One motivation for spotting research aspects automatically is to help search and information extraction (Teufel et al., 1999). We have teamed up with the group who created COVIDSeer[4] to explore the possible uses of CODA-19 in such systems.

What other types of biomedical annotations can be crowdsourced? Many prior works that used crowd workers to annotate medical documents (Khare et al., 2016) focused on images (Heim et al., 2018) or named entities (*e.g.,* medical terms (Mohan and Li, 2019), disease (Good et al., 2014), or medicine (Abaho et al., 2019).) We will explore what other types of annotations can be created using non-expert workers.

Acknowledgments

This project is supported by the Huck Institutes of the Life Sciences' Coronavirus Research Seed Fund (CRSF) at Penn State University and the College of IST COVID-19 Seed Fund at Penn State University. We thank the crowd workers for participating in this project and providing useful feedback. We thank VoiceBunny Inc. for granting a 20% discount for the voiceover for the worker tutorial video to support projects relevant to COVID-19. We also thank Tiffany Knearem, Shih-Hong (Alan) Huang, Joseph Chee Chang, and Frank Ritter for the great discussion and useful feedback.

[4] CovidSeer: https://covidseer.ist.psu.edu/

References

Micheal Abaho, Danushka Bollegala, Paula Williamson, and Susanna Dodd. 2019. Correcting crowdsourced annotations to improve detection of outcome types in evidence based medicine. In *CEUR Workshop Proceedings*, volume 2429, pages 1–5.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. *arXiv preprint arXiv:1704.02853*.

Soumya Banerjee, Debarshi Kumar Sanyal, Samiran Chattopadhyay, Plaban Kumar Bhowmick, and Parthapratim Das. 2020. Segmenting scientific abstracts into discourse categories: A deep learning-based approach for sparse labeled data. *arXiv preprint arXiv:2005.05414*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. Solvent: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–21.

Giovanni Colavizza, Rodrigo Costas, Vincent A. Traag, Nees Jan van Eck, Thed van Leeuwen, and Ludo Waltman. 2020. A scientometric overview of cord-19. *bioRxiv*.

Pradeep Dasigi, Gully APC Burns, Eduard Hovy, and Anita de Waard. 2017. Experiment segmentation in scientific discourse as clause-level structured prediction using recurrent neural networks. *arXiv preprint arXiv:1702.05398*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haifa Zargayouna, and Thierry Charnois. 2018. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688.

Benjamin M Good, Max Nanis, Chunlei Wu, and Andrew I Su. 2014. Microtask crowdsourcing for disease mention annotation in pubmed abstracts. In *Pacific Symposium on Biocomputing Co-Chairs*, pages 282–293. World Scientific.

James Hartley. 2004. Current findings from research on structured abstracts. *Journal of the Medical Library Association*, 92(3):368.

Eric Heim, Tobias Roß, Alexander Seitel, Keno März, Bram Stieltjes, Matthias Eisenmann, Johannes Lebert, Jasmin Metzger, Gregor Sommer, Alexander W Sauter, et al. 2018. Large-scale medical image annotation with crowd-powered algorithms. *Journal of Medical Imaging*, 5(3):034002.

Hen-Hsen Huang and Hsin-Hsi Chen. 2017. Disa: A scientific writing advisor with deep information structure analysis. In *IJCAI*, pages 5229–5231.

Natalia B Hubbs, Mareena M Whisby-Pitts, and Jonathan L McMurry. 2019. Kinetic analysis of bacteriophage sf6 binding to outer membrane protein a using whole virions. *bioRxiv*, page 509141.

Ritu Khare, Benjamin M Good, Robert Leaman, Andrew I Su, and Zhiyong Lu. 2016. Crowdsourcing in biomedicine: challenges and opportunities. *Briefings in bioinformatics*, 17(1):23–32.

Halil Kilicoglu. 2018. Biomedical text mining for research rigor and integrity: tasks, challenges, directions. *Briefings in bioinformatics*, 19(6):1400–1414.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International journal of medical informatics*, 75(6):468–487.

Sunil Mohan and Donghui Li. 2019. Medmentions: a large biomedical corpus annotated with umls concepts. *arXiv preprint arXiv:1902.09476*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830.

Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The acl rd-tec 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1862–1868.

Eva Siegenthaler, Yves Bochud, Per Bergamin, and Pascal Wurtz. 2012. Reading on lcd vs e-ink displays: effects on fatigue and visual strain. *Ophthalmic and Physiological Optics*, 32(5):367–374.

Simone Teufel et al. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, Citeseer.

A. de Waard, S. Buckingham Shum, A. Carusi, J. Park, M. Samwald, and Á. Sándor. 2009. Hypotheses, evidence and relationships: The hyper approach for representing scientific knowledge claims. In *Proceedings 8th International Semantic Web Conference, Workshop on Semantic Web Applications in Scientific Discourse. Lecture Notes in Computer Science, Springer Verlag: Berlin*.

Anita de Waard and Henk Pander Maat. 2012. Verb form indicates discourse segment type in biological research papers: Experimental evidence. *Journal of English for Academic Purposes*, 11(4):357–366.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Michael Kinney, Ziyang Liu, William. Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*, abs/2004.10706.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. Paperrobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1980–1991.

W John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC bioinformatics*, 7(1):356.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.